Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 91 (2016) 971 – 977

Procedia Computer Science

Information Technology and Quantitative Management (ITQM 2016)

A Multi-level Traceability System based on GraphLab

Wei Teng[a], Bo Mao[a,] *, Jie Cao[a]

[a]College of Information Engineering, Collaborative Innovation Center for Modern Grain Circulation and Safety, Jiangsu Key Laboratory of Modern Logistics, Nanjing University of Finance & Economics, Nanjing, China

Abstract

In this paper a multi-level traceability system is proposed. First traceability information including logistics and environment data is collected from each link of the supply chain. Then the data is stored locally with GraphLab, abstracted and sent to the higher level traceability data center. In the demonstration, we simulate the traceability data collection process and deploy a GraphLab server to fulfill the traceability task such as consumer distribution and product tracking. The experimental results indicate that GraphLab can supply efficient development platform and visualize the traceability information.

Peer-review under responsibility of the Organizing Committee of ITQM 2016

Keywords: Traceaility System; Graph mining; Multi-level data

1. Introduction

Traceability is essential in food security especially in China where the supervision is not as soundly as in Western countries. Many accidents related with food or drug security have been reported every year. For example the recent China vaccine scandal that exposes weaknesses in how vaccines are distributed across the country. It is again reminder us that the traceability system is necessary for essential products such as food, drugs, agriculture materials and so on. On the other hand, the traceability system should not only tell us where the product comes from, but also the environment information of the whole supply chain. The vaccine scandal is mainly coursed by the fail of temperature control during the storage and transportation.

In 2015, the Food and Drug Administration of China lunched a project to create a national wised traceability framework for food. Meanwhile, All China Federation of Supply and Marketing Cooperatives starts a task to develop a traceability and anti-fake technologies for agriculture materials. As a research center for grain storage and informationization, we participate in both projects to develop the traceability infrastructure for China. According to our study, the information required for traceability is extremely large which covers raw material

* Corresponding author. Tel.: +86-25-83493900; fax: +86-25-83493900.
E-mail address: maoboo@gmail.com.

Peer-review under responsibility of the Organizing Committee of ITQM 2016
doi:10.1016/j.procs.2016.07.126

supply, process of production, storage, transportation, sales, and the consumer. All traceability and environment details including temperature, humility, light, CO2, PM2.5, surveillance video, and et al. about these links are considered necessary in the traceability system. Take the huge number of producer and products, the data volume will be big.

Currently, the official administration has asked the producer to report their raw material and production information in batch and these information is required to be persevered for one year. The existing system constructed with Oracle database on high performance server. However, along with the million updates each day, the system suffers from frequently updating in storage and continuously decreasing in performance. It is necessary to design a new structure for the traceability system that will contains more data in various sources.

In this paper, we presented a distributed traceability data management structure using GraphLab and tested it with the generated traceability data. The rest of the paper is organized as follows. Related work is introduced in Section 2; Section 3 presents the proposed traceability framework; Experimental results are explained in detail in Section 4; finally Section 5 concludes the paper.

2. Related Work

Traceability framework. The traceability for food and agriculture product is widely studied over the world. To implement the traceability system, logistic information should be recorded through the whole supply chain. Currently RFID is the mainly used method for logistic data [1]. Meanwhile, stable isotope [2] and DNA barcode [3] are introduced for more accurate traceability data collection. Besides the logistic data, the environment information related white the whole supply chain is also related to the quality of the target product and should be recorded [4]. In cold meat chain, temperature is monitored in real time for the traceability purpose [5]. Humidity is also recorded for food quality in cold storage system [6]. Video surveillance data is also used for egg [7] and grain [8] traceability system. Along with the development of Internet of Things, increasing volume of data related to traceability supply chain will be automatically generated. To integrate environment data with logistic data and to perform high efficiency analysis, a big graph based mining framework is required.

Graph mining. Graph is a common structure in many applications. For example World Wide Web[9], computer networks[10], social networks[11], supply chain[12] can be described as graphs. The nodes and edges in these graphs are usually quite large, which leads the mining and analysis of the big graph a difficult task for the traditional standalone computer structure. To deal with big graphs, a parallel or distributed framework is required. MapReduce[13] or its open source implementation Hadoop is one of the most widely used parallel computation framework. It has been used to build several big graph analysis applications [14]. Because of the limitation of Hadoop such as lack of random graph data access, there is a bottleneck for graph mining with Hadoop. To improve the graph analysis capacity of Hadoop, a scalable and general graph management and mining system, GBase [15] is proposed. Also a fast method for big graph mining based on HBase [16] is introduced. However these methods are based on Hadoop structure and it is difficult to integrate existing data source and require a complete system rebuild. GraphLab [17] is a new parallel framework to deal with big graph proposed by researchers from CMU. It not only support the common machine learning algorithms but also is compatible with different data sources such as file, SQL database, CSV, Hadoop file, Spark RDD and et al. In this paper, we deploy GraphLab in the traceability analysis to provide a new method for food or agriculture material.

3. Methodology

In this paper, GraphLab is explored to deal with traceability data from multiple sources in different format. First the overall framework of the system is introduced. Then we explain the data integration strategy for

multiple sources. Based on the integrated data, basic traceability analysis methods such as search, statistics, and outlier detection are studied using GraphLab. Finally, visualization of traceability results is implemented with GraphLab.

3.1. Overall framework

In the real situation, each producer can input raw materials and output products. Multiple producers and customers combined with transportation form the product supply chain. Usually each link in the supply chain are separately managed by different manufacturers. Therefore, traceability data should be collected in each link and merged together for the overall analysis. However, considering the large volume of logistic and environment data related to the traceability analysis, it is difficult and costly to transfer and maintain all of the traceability data in a single data center. In this paper, we presented following distributed traceability data analysis framework based on GraphLab as shown in Fig. 1.

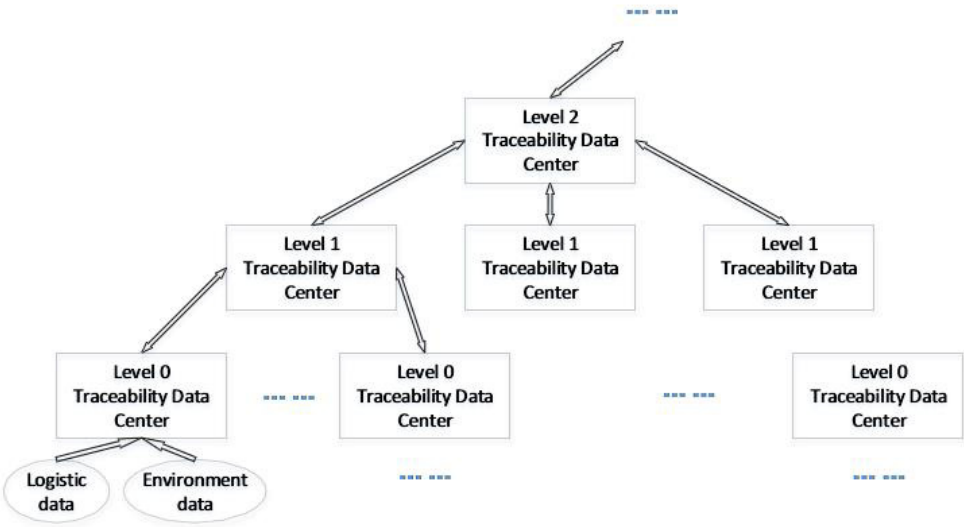

Fig. 1 GraphLab based Multi-level Traceability Framework

In this framework, traceability data is collected and managed locally. The abstractions or summarizes of the collected logistic and environment data about the processing or storage are transform to higher level traceability data note. The overall structure of the system is tree based and can be extended into any scales. For example, each factory can have the traceability nodes connected with the city level nodes that further link to the province level and finally the national level. In each level, summarized data from lower levels are stored and analyzed to generate the abstraction of current level information and send it to the higher level. Meanwhile, the higher level can require the detail information about the traceability from lower level and send summarize data to lower level if it is necessary.

3.2. Data integration

The collected traceability data includes logistic and environment information. Each node receives data from lower layer nodes, integrates the data and send the abstracted data to upper layer node. The data integration contains three main steps association, mining and abstraction.

In the association step, data from different lower layer nodes are connected according to the relationship of logistic chain and timestamp. To implement the association, we convert the supply chain into a SGraph structure, in which products are saved as vertex and actions such as storage, processing or transportation as edges in GraphLab. In each edge, timestamps are recorded when the action starts, ends or something happens. In the vertex, the quality and quantity attributes of the product is stored. GraphLab supports various data sources such as S3, ODBC, JSON, CSV, HDFS and etc. Therefore, it is easy to import data from different companies or organizations in the supply chain of food or agriculture related products.

In mining step, GraphLab supplies a powerful algorithm library that covers the main stream data mining fields such as classification, clustering, regression and so on. It is convenient to employ these providing methods for different traceability purposes for example searching, exception detection, influence evaluation et al.

In abstraction step, collected data are filtered and summarized to describe the basic situation of the current node that will be uploaded to the higher layer in the system. Besides the daily operation abstraction, if emergency situation related to the product quality happens such as pollution source detected or temperature abnormity, the current node will upload the situation to higher level nodes or directly to the emergency processing nodes.

3.3. Traceability Analysis

Multi-level analysis

In the proposed traceability data integration framework, each node is capable of providing the traceability information based on its collected date. And it is required the top layer node to provide the traceability covers the whole supply chain. To combine the ability of entire traceability system, we proposed a multi-level detail traceability analysis.

In fact the traceability information is similar with fractal geometry in which each part can be further separated into smaller parts. Each link in the supply chain can be divided into sub sections for example the meat processing contains slaughtering, partition, cleaning and storage, and these sections can be further separated into more detailed parts. Meanwhile it is nature for the user to check the traceability information zoom in and out in different levels as digital map. Therefore, it is necessary to organize and analysis the traceability data in multi-level of details.

Deep learning methods

Besides multi-level analysis, GraphLab also support deep learning framework. This is useful to create the outlier detection model and perform the semantic classification especially for video surveillance data. GraphLab has included the mainstream deep learning model into its library such as CNN, RNN and et al. We can contracture and train the deep learning model with the proposed methods. An easier way to use the deep learning function of GraphLab is to load the pre-trained models directly to analysis the data.

As a powerful analysis tool, GraphLab provides quite many useful algorithm to deal with the traceability data which makes the researchers focus on the business instead of programming details.

4. Experimental results

4.1. Simulated data generation

To generate the traceability data, we define a supply chain with three types of elements: producer, transporter and consumer. Each producer takes into n input material and output m products in s hours. The transporter transits products from producer pi to pj. The consumer buys the final products and ends the supply

chain. In each element, we can generate the logistic and environment data (timestamp, temperature, humidity and surveillance video) as the traceability record and save the data for further analysis. Each element can record its own traceability information, receive and send data into the traceability center where GraphLab is deployed.

In the simulation, each element (producer, transporter or consumer) is running as a standalone threads and communicates with others through queue structure. For example, the producer only start to produce the product if its storage queue contains all raw materials it needs. The transporter will start if it receive enough products.

Based on the proposed traceability simulation system, we can easily test the efficiency and accuracy of the analysis system. The simulated supply chain is visualized as follows, in which green nodes are producers, gray nodes are transporters and red ones are consumers.

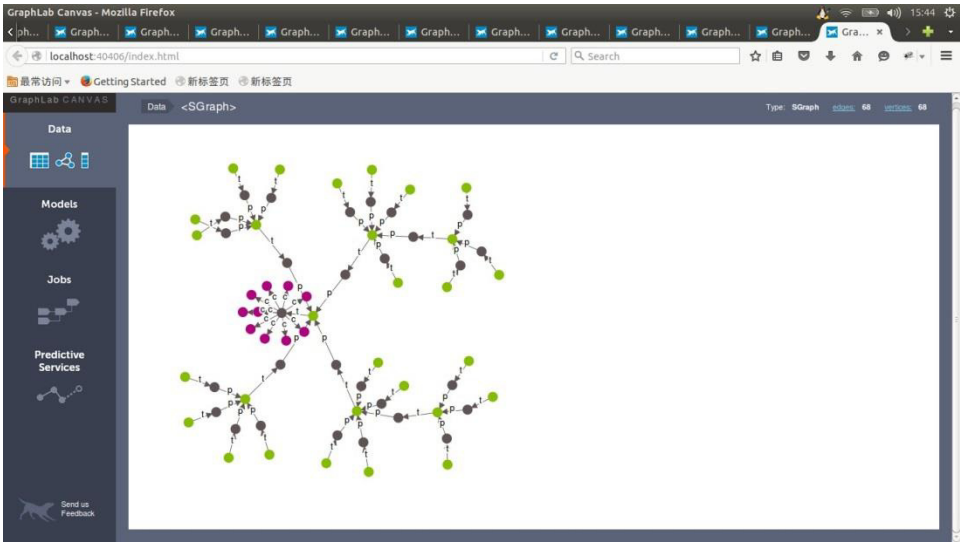

Fig 2. Simulated supply chain

4.2. GraphLab platform

To test the GraphLab based traceability analysis method, we build a server with GraphLab-Create 1.8.5 with Ubuntu 14.04 LTS operation system. The server is a common laptop with 4G RAM and Core i3 CPU. The system is deployed with conda development environment using python 2.7 as the GraphLab official site suggested.

4.3. Traceability visualization

Based on the proposed framework, we have implemented a demo traceability analysis framework that can supply the product trace back and raw material distribution analysis. As shown in Fig. 3, the consumer (red node) can find out how its food is produced and transformed. Meanwhile, we can analysis which consumers receive the products that contain a certain raw material or come from a certain producer as given in Fig. 4.

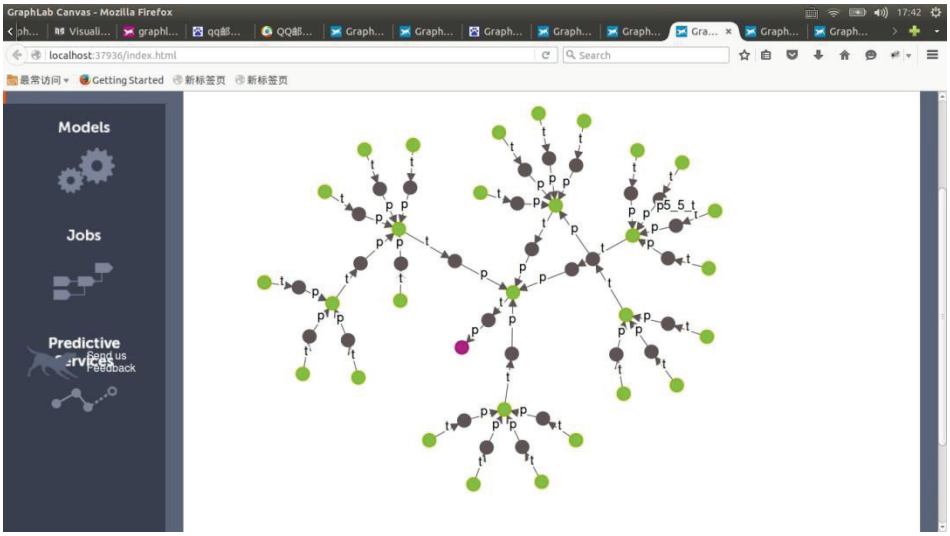

Fig 3. Trace back of the product

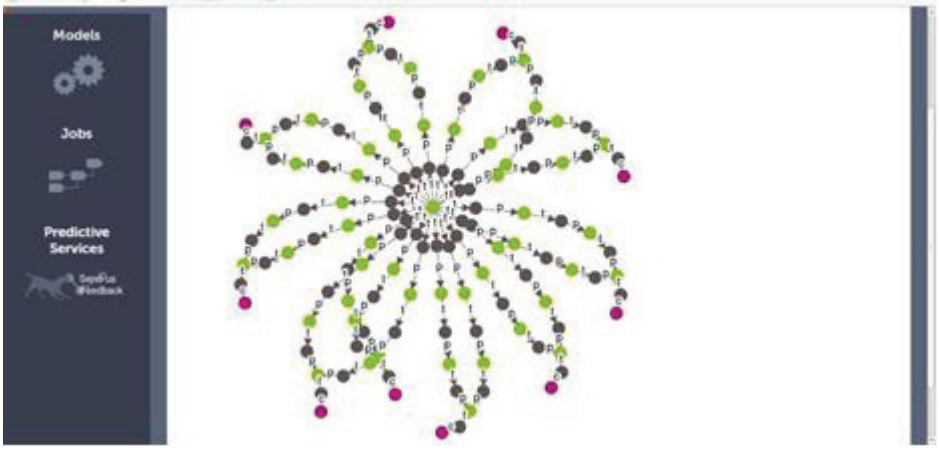

Fig. 4. Raw material distribution or user influence

5. Conclusion

Traceability is essential for the safely and anti-fake of food and agriculture related product. Considering the complexity of supply chain, we propose a multi-level traceability data analysis framework by collect and integrate both logistic and environment information. The traceability data is stored in GraphLab a distributed

Graph analysis tool to perform the mining and visualization of the data. We implement a single node analysis system for the simulated traceability data. The tests show that GraphLab can efficiently process the traceability data and supply a potential for further complex mining such as deep learning. In the future, we will

implement the proposed system in the real food or agriculture product traceability situation, and perform deep learning based outlier detection.

Acknowledgements

This work was supported by the Natural Science Foundation of Jiangsu (Grant No. BK20151551), National Key Technologies R&D Program of China (Grant No.2015BAD18B02 and 2015BAK36B02), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 1402120C), National Center for International Joint Research on E-Business Information Processing under Grant 2013B01035, China Special Fund for Grain-scientific Research in the Public Interest (201513004) ,Independent Innovation for Agricultural Science of Jiangsu cx(15)1051 and the project of the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Nanjing University of Finance and Economics.

References

[1] Piramuthu S, Farahani P, Grunow M. RFID-generated traceability for contaminated product recall in perishable food supply networks. European Journal of Operational Research 2013; 255(3): 253-262.
[2] Portarena S, Gavrichkova O, Lauteri M , Brugnoli E 2014. Authentication and traceability of Italian extra-virgin olive oils by means of stable isotopes techniques. Food chemistry;164:12-16.
[3] Galimberti A, De Mattia F, Losa A , Bruni I, Federici S, Casiraghi M , Labra M. DNA: barcoding as a new tool for food traceability. Food Research International 2013; 50(1): 55-63.
[4] Aung M M , Chang Y S. Traceability in a food supply chain: Safety and quality perspectives. Food control 2014; 39:172-184.
[5] Thakur M , Forå s,E. EPCIS based online temperature monitoring and traceability in a cold meat chain. Computers and Electronics in Agriculture 2015; 117:22-30.
[6] Kim W R, Aung M M , Chang Y S , Makatsoris C. Freshness Gauge based cold storage management: A method for adjusting temperature and humidity levels for food quality. Food Control 2015; 47:510-519.
[7] Liu F , Wang Y, Jia Y, Hu S, Tu L ,Tang C. The egg traceability system based on the video capture and wireless networking technology. International Journal of Sensor Networks 2015;17(4):211-216.
[8] Mao B, He J, Cao J, Gao W , Pan D. Food Traceability System Based on 3D City Models and Deep Learning. Annals of Data Science 2016; 1-12.
[9] Andrei B, Ravi K , Farzin M, Prabhakar R, Sridhar R, Raymie S, Andrew T, Janet W. Graph structure in the web. Computer Network 2000; 33 (1–6) :309–320.
[10] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology.SIGCOMM Comput.Commun.Rev 1999; 29 (4):251–262.
[11] Ellison N B, Steinfield C, Lampe C. The benefits of facebook friends: social capital and college students use of online social network sites. J. Comput.-Mediated Commun.2007;12 (4):1143–1168.
[12] Tan K H, Zhan Y Z, Ji G J . Ye F, Chang C. Harvesting big data to enhance supply chain innovation capabilities: An analytic infrastructure based on deduction graph. International Journal of Production Economics 2015;165:223-233.
[13] Dean J, Ghemawat S. Mapreduce: simplified data processing on large clusters 2004. In: 6th Symposium on Operating System Design and Implementation (OSDI 2004), San Francisco, California, USA, p.137–150.
[14] Kang U, Tsourakakis C E, Appel A P, Faloutsos C, Leskovec J. HADI: mining radii of large graphs. TKDD 2011; 5 (2):8.
[15] Kang U, Tong H, Sun J, Lin C, Faloutsos C. Gbase: an efficient analysis platform for large graphs.VLDB 2012; 21 (5) :637–650.
[16] Lee H, Shao B, Kang U. Fast graph mining with HBase. Information Sciences 2015; 315:56-66.
[17] Low Y, Bickson D, Gonzalez J, Guestrin C, Kyrola A ,Hellerstein J M . Distributed GraphLab: a framework for machine learning and data mining in the cloud. Proceedings of the VLDB Endowment 2012; 5(8):716-727.